

Estimation of Relative Survival Based on Cancer Registry Data

Olaf Schoffer^{*1}, Antje Niedostatek², Stefanie J. Klug^{1,2}

¹Cancer Epidemiology, University Cancer Center, TU Dresden, Fetscherstrasse 74, 01307 Dresden, Germany

²Regional Clinical Cancer Registry Dresden, Fetscherstrasse 74, 01307 Dresden, Germany

^{*}olaf.schoffer@uniklinikum-dresden.de; ²antje.niedostatek@uniklinikum-dresden.de;

^{1,2}stefanie.klug@uniklinikum-dresden.de

Abstract

Relative survival rates are suitable for the comparison of survival information between patient groups with different age structures. They are well established as a common method for analysis of cancer registry data. In the following, the method for the estimation of the relative survival and the central components of the implementation of this method with SAS are presented. Examples of results of the described program are presented and the program is compared with a selection of other technical solutions.

Keywords

Relative Survival; Cancer Registry; Natural Mortality; Kaplan-Meier Estimate; Nelson-Aalen Estimate; Ederer II; Pohar Perme Method

Introduction

The estimation of the survival function is an essential focus of epidemiological research and frequently used in analyses of cancer registry data. Besides the particular disease, age of the observed patient group extensively influences survival time. Relative survival rates are used to decompose the observed mortality into a population-based hazard and an excess hazard caused by the particular disease (net survival). Thus, the influence of the age structure is eliminated.

Methods

The most established method for the estimation of survival functions is the Kaplan-Meier estimation. Here, the survival function for events at time $t_1 < t_2 < \dots < t_k$ is estimated to

$$\hat{S}(t) = \begin{cases} 1, & t = 0 \\ \prod_{i=1}^j \left(1 - \frac{d_i}{n_i}\right), & t_j \leq t < t_{j+1} \leq t_k \\ \prod_{i=1}^k \left(1 - \frac{d_i}{n_i}\right), & t \geq t_k \end{cases}$$

whereas n_i indicates the number of individuals under risk at time t_i and d_i the number of events at this point in time (Allison 2010). An individual is under risk at a specific point in time if the person has not experienced an event (death) or was not censored (lost-to-follow-up).

The selection of the population occurs mostly by providing data for a certain time period where an individual is part of the population (e.g. date of diagnosis in the years 2000 to 2009). This procedure is called cohort approach.

Brenner and Gefeller (1996) introduced the period approach. In comparison to the cohort approach, the period approach only uses events and censoring, which occur in a predefined time interval. The population under risk is only defined for this specific time interval.

The Life Table estimation (Cutler and Ederer 1958) is also commonly used, but differs from the Kaplan-Meier estimation. On one hand, the time points t_i do not depict the events' time points but the borders of a priori chosen (equidistant) intervals. On the other hand, n_i is replaced by $n_i - (c_i/2)$ in the above mentioned equation, in which c_i indicates the number of censoring within the interval $[t_{i-1}; t_i)$. Since the intervals are often selected wider than the possible time interval (for example years versus exact day), an impreciseness has to be accepted here.

Another method is the Nelson-Aalen estimator for the cumulative hazard function. This is defined as

$$\tilde{H}(t) = \begin{cases} 1, & t = 0 \\ \sum_{i=1}^j \frac{d_i}{n_i}, & t_j \leq t < t_{j+1} \leq t_k \\ \sum_{i=1}^k \frac{d_i}{n_i}, & t \geq t_k \end{cases}$$

whereas t_i , d_i and n_i are defined according to the

Kaplan-Meier estimator. The continuous version of the Nelson-Aalen estimator is

$$\tilde{H}(t) = \int_0^t \frac{dD(u)}{N(u)}$$

with $D(t) = \sum_{i=1}^n D_i(t)$ where $D_i(t)$ is a counting process and with $N(t) = \sum_{i=1}^n N_i(t)$ where $N_i(t)$ is the at-risk process for an individual i ($i = 1, \dots, n$) as defined in Pohar Perme, Stare, and Estève (2012). Hence, the Breslow estimator for the survival rate is

$$\tilde{S}(t) = e^{-\tilde{H}(t)}$$

with an estimated standard error (Greenwood formula (Kalbfleisch and Prentice 1980))

$$\hat{\sigma}(\tilde{S}(t_j)) = \tilde{S}(t_j) \sqrt{\sum_{i=1}^j \frac{d_i}{n_i(n_i - d_i)}}.$$

A suitable confidence interval with boundaries within the given interval $[0;1]$ can be estimated using a transformation of $\tilde{S}(t)$ according to Borgan and Liestol (1990).

The Breslow estimator is asymptotically equivalent to the Kaplan-Meier estimator and can be derived by the Martingale theory via counting process (Aalen 1978). The approximation improves with increase of population under risk in relation to the number of events per points in time.

The advantages and disadvantages of both methods are discussed in Colosimo and colleagues (2002). No advantages of one method compared to the other can be found. However, distinct differences appear within small samples.

However, for a number of applications, including the analysis of cancer registry data, the estimation of observed survival should be extended as follows. Information on mortality independent from age and sex structures of the observed population is necessary to allow for comparisons of mortality rates between countries. For this, the concept of relative survival should be used. The basic idea of this concept is to divide the total mortality into natural mortality and excess mortality (Koch 2001). The natural mortality at time t describes the proportional mortality of the standard population which is identical regarding sex and age to the observed population under risk. Then the mortality of the standard population can be derived from sex and age specific mortality tables of the official statistics, for example the period mortality tables of the German Federal Statistical Office (Statistisches Bundesamt 2011). The remaining excess mortality after elimination of the natural mortality

from the total mortality describes the additional mortality due to a specific disease of the observed population.

The Ederer II procedure (Ederer and Heise 1959) is used for the estimation of natural mortality. The derivation of the relative survival is described in the following.

1. Replication of information of the population for each event time point (i.e. Cartesian product of n_0 individuals und k event's time points).
2. Restriction of the retrieved data basis for each event time point on the respective population under risk.
3. Determination of the age of each individual at each event time point.
4. Assignment of the natural mortality (year, age and sex specific) for each individual at each event time point.
5. Per event time point: estimation of the mean over the individual natural mortalities of all individuals under risk at the respective event time point.
6. Weighing of the mean mortality at time point t_i . As weight, the time past the previous event time point ($t_i - t_{i-1}$) is used.
7. Accumulation of the weighted natural mortality over all time points t_i to: $H^*(t)$ ($t_i \leq t < t_{i+1}$).

The continuous version of the natural mortality is

$$H^*(t) = \int_0^t \frac{\sum_{i=1}^n N_i(u) dH_{P_i}(u)}{N(u)}$$

whereas $H_{P_i}(t)$ represents the natural mortality for an individual i at time t .

Analogous to the derivation of the Breslow estimator from the Nelson-Aalen estimator, the expected survival rate from the natural mortality is estimated as follows

$$S^*(t) = e^{-H^*(t)}.$$

Thereby, the relative survival rate is estimated as

$$S_{rel}(t) = \frac{\tilde{S}(t)}{S^*(t)}.$$

Since the natural mortality is estimated from the total standard population and, therefore, should not include error of estimation, the estimated standard error is, according to the additive hazard model, as follows

$$\hat{\sigma}(S_{rel}(t)) = \hat{\sigma}(\tilde{S}(t)).$$

A very promising alternative to the Ederer II method is the Pohar Perme method (Pohar Perme, Stare, and Estève 2012), which provides an unbiased estimation of the net survival. In this method, the terms $D_i(t)$ and $N_i(t)$ in the continuous notations of the Nelson-Aalen estimator and the natural mortality are replaced by their weighted counterparts $D_i^w(t) = D_i(t)/S_{p_i}(t)$ and $N_i^w(t) = N_i(t)/S_{p_i}(t)$, whereas $S_{p_i}(t)$ is the expected survival rate for an individual i at time t . Hence, relative survival rate is

$$S_{rel}^{PPM}(t) = \exp \left\{ \int_0^t \frac{\sum_{i=1}^n N_i^w(u) dH_{p_i}(u)}{N^w(u)} - \int_0^t \frac{dD^w(u)}{N^w(u)} \right\}.$$

The relative survival rate describes the proportion of survivors based on the part of the population that would have survived even in the age and sex identical standard population.

For example, if the Breslow estimator calculates a five-year total survival of 60%, whereas the reference population in the same time frame indicates an expected survival of 90%, the relative survival is 66.7% ($=60\% / 0.9$). If the total survival of the observed population is larger than the expected survival in the standard reference population, the relative survival can be above 100%.

Presentation of Results

The described methods are implemented in a SAS program that is briefly described in the appendix. It produces graphics and tabular overviews for the estimated total, expected and relative survival. This will in part be presented based on an example of data from clinical cancer registries (CCR). In the frame of an initiative of the Working Group of German Clinical Cancer Registries (ADT), 22 population-based and 4 institution-based CCR provided, for an evaluation of data of malignant melanoma (ICD-10: C43 and D03), more than 45 000 primary cases notified within the time period 2000-2009. Among them, more than 30 000 cases were part of the performed survival analysis (Breslow estimator with Ederer II method and cohort approach). Only population-based registries with sound information on the vital status of the cancer patients were included in the survival analysis. Case data that were based on multiple notifications of the same patient were excluded from the analysis. Due to the large amount of data connected with the above mentioned asymptotic

equivalence, the Breslow estimator is a close approximation to the Kaplan Meier estimator.

Figure 1 shows the total, the expected and the relative 5-year survival. For the relative survival, the 95% confidence interval is presented.

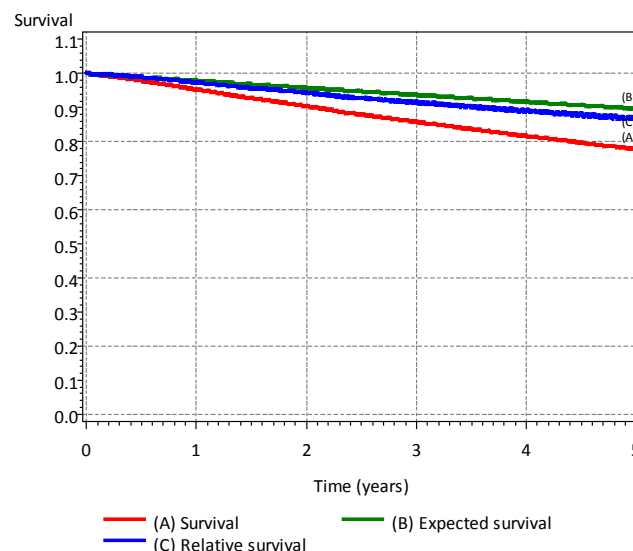


FIG. 1 5-YEAR SURVIVAL (TOTAL, EXPECTED, RELATIVE) OF MALIGNANT MELANOMA

The 5-year relative survival rate using Ederer II method and cohort approach is 86.8%. Using the Pohar Perme method, the 5-year relative survival rate is 86.3%, which is within the confidence bound of the relative survival using Ederer II. Hence, the bias of the Ederer II method is marginal. Using period approach with a period range covering the years 2006 to 2009, the 5-year relative survival rate is 85.5%.

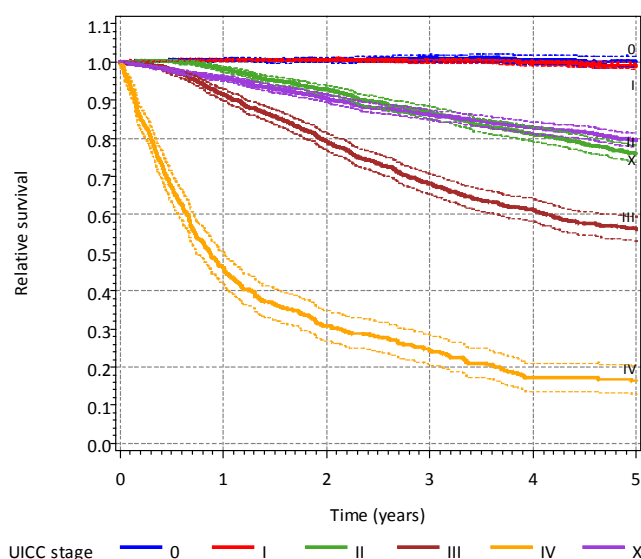


FIG. 2 5-YEAR RELATIVE SURVIVAL OF MALIGNANT MELANOMA STRATIFIED BY UICC STAGE

Figure 2 shows the relative survival with the 95% confidence intervals stratified by UICC stages

(Wittekind and International Union Against Cancer (UICC) 2003). For the stages 0 and 1 the relative survival is approximately 100%. For higher stages, a decrease in survival compared to the standard population can be seen. The severity of disease has considerable consequences for the survival of the patients, even though the information of the patients' age structure was removed.

Table 1 quantifies the relation between relative, expected and total survival shown in Figure 1. Comparing the total survival between both sexes 5 years after diagnosis, women have a considerably better perspective with a survival rate of 81.5% than men with a survival rate of 74.2%. However, a group of the standard population identical to the age of the observed population would also have a better survival for women (expected survival rate: 91.4%) than for men (expected survival rate: 87.9%). It can be questioned whether the better survival of women versus men with a malignant melanoma is solely attributable to the already lower life expectancy of men or whether one can still see this difference between both sexes after the removal of this effect. After the removal of expected survival, an advantage of (relative) survival in women remains.

TABLE 1 SURVIVAL (TOTAL, EXPECTED, RELATIVE) AFTER DIAGNOSIS
"MALIGNANT MELANOMA" STRATIFIED BY SEX

Sex	Year (after diagnosis)	Survival	Expected Survival	Relative Survival	Relative Survival 95%-LCL	Relative Survival 95%-UCL
Men	1	94.6%	97.5%	97.0%	96.6%	97.4%
	2	88.8%	95.1%	93.4%	92.8%	94.0%
	3	83.3%	92.7%	89.9%	89.0%	90.6%
	4	78.3%	90.3%	86.8%	85.8%	87.7%
	5	74.2%	87.9%	84.4%	83.3%	85.4%
Women	1	96.1%	98.3%	97.8%	97.4%	98.1%
	2	92.1%	96.5%	95.4%	94.8%	95.9%
	3	88.4%	94.8%	93.2%	92.5%	93.8%
	4	85.0%	93.1%	91.3%	90.5%	92.1%
	5	81.5%	91.4%	89.2%	88.2%	90.1%

Comparison with other Solutions

Table 2 compares the described implementation (Koch 2001) with a selection of other established solutions in regard to the used software and available methods.

An analysis with the program SURVSOFT (Geiss, Meyer, Radespiel-Troger, and Gefeller 2009) analogous to table 1 reveals almost identical results for the percentage and the confidence intervals of the relative survival. Deviation from the 5-year survival

occurs only in the second decimal place of the percentage values. Other solutions result in slightly larger deviations, due to different data accurateness or method. The conclusions remain, however, comparable.

Besides the described differences and the similarities of the different solutions, distinct differences partly remain in regard to the presentation of results and the handling, which can, however, not be described in more detail here.

TABLE 2 OVERVIEW OF EXAMPLE SOLUTIONS TO ESTIMATE THE RELATIVE SURVIVAL

Solution	Software	Method		
		Survival	Expected Survival	Relative Survival
Extension of RSURV (Koch 2001)	SAS	Life Table, Kaplan-Meier & Breslow	Ederer II & Pohar Perme*	Cohort & Period Approach*
SURVSOFT (Geiss et al. 2009)	Own Development	Life Table & Kaplan-Meier	Ederer II & Hakulinen (Hakulinen 1982)	Cohort & Period Approach
Programs of P. Dickman (Dickman 2004)	SAS / Stata	Life Table	Ederer II & Pohar Perme	Cohort & Period Approach
period(R/H) (Brenner, Gefeller, and Hakulinen 2004)	SAS / R	Life Table	Ederer II & Hakulinen	Period** Approach

*The simultaneous use of the Pohar Perme method with the period approach is not yet implemented

**Cohort estimation can be performed by choosing a period range that spans the entire range of data

Conclusions

With the presented program, a flexible and comparable solution in regard to the results of established programs for the estimation of relative survival rates is available. Small differences in the results occur due to the different methods in regard to, for example, the expected and the total survival, the selection of the observed population, the accurateness of the time axis and the life table used.

Due to the reduced complexity of syntax and macro functionality, the program can be adapted more flexibly and offers the whole spectrum of methods which are available with the procedure LIFETEST Since SAS 9.22. The produced results can be reported in different file formats via ODS and SAS/GRAPH.

During programming of the syntax, an optimal program flow was considered. Special focus was set

on the reduction of the data base already during the establishment of the Cartesian product and the consequent exploitation of improvements until SAS 9.22 (LIFETEST; ODS; SQL).

Appendix

Implementation in SAS

The following implementation in SAS is an advancement of the program RSURV (Koch 2001). However, only a few essential steps and not the complete program are presented here. The combination of the available programming elements with the aim to estimate the relative survival has, to our knowledge, not been described elsewhere.

Macro-Functionality

The presented estimation was embedded into a macro that is defined by keyword parameters: %macro rsurv_s(ds=, strat=, methd=, net=, approach=, nzus=, tit2=);

The macro variables for the names of the base data set (&ds), the stratifying variable (&strat), the used method for estimating observed survival (&methd), the used method for calculating net survival (&net), and the chosen cohort or period approach (&approach) will be needed in the following. The method for estimating observed survival can be chosen according to the available options in PROC LIFETEST. For the base data set, a data set with the following variables is expected: patid (unique patient ID), ddate (date of diagnosis), bdate (date of birth), time (number of days between date of diagnosis and date of event or censoring), status (0 $\hat{=}$ censored observation, 1 $\hat{=}$ observation with event), sex (1 $\hat{=}$ male, 2 $\hat{=}$ female). Further parameters could be a text for the subtitle in the output (&tit2) and a data name add on (&nzus).

Besides those, predefined macro variables are used within the macro. Those refer to the output format (&channel), for example PDF or RTF, the maximal depicted observation time (&maxtime) as well as the indication of the path for data output (&path).

A further description of the used macro functionality can be found elsewhere (Krämer, Schoffer, and Tschiersch 2008) and (SAS Institute Inc. 2009).

PROC LIFETEST with ODS

A central component of the estimation of relative survival rates is the estimation of the total survival (by Life Table, Kaplan Meier or Breslow estimator). This

has been implemented for the cohort approach with the procedure LIFETEST (SAS Institute Inc. 2010) in SAS since version 9.22.

For the period approach and the Pohar Perme method, the total survival has to be estimated without the procedure LIFETEST.

The procedure request used in the program and some options are briefly described.

```
PROC LIFETEST DATA=&ds
  METHOD=BRESLOW NELSON ATRISK
  INTERVALS=(0 TO &maxtime BY 1)
  CONFTYPE=LOGLOG
  OUTSURV=&ds._cl;
  TIME time*status(0);
  STRATA &strat;
RUN;
```

In SAS versions 9.2 and 9.22, the following options for the LIFETEST procedure are newly introduced or extended. The ATRISK option is used to output the number of individuals under risk. With CONFTYPE= the transformation to estimate the confidence intervals is indicated. The METHOD= option now allows the selection of the Breslow estimator to estimate the total survival in addition to the earlier available methods Kaplan-Meier and Life Table. The NELSON option is used to output the Nelson-Aalen estimator.

The result output in the program is operated by ODS. The indication for the output channel is given by the macro variable &channel, which is defined at the begin of the program. In addition to the text and graphic output, ODS is also used to retrieve estimation results as SAS data set. This data set is the basis for the subsequent estimation of the relative survival and is later linked to the data set which includes the confidence estimation for the total survival. For this, PROC LIFETEST is enclosed by the following ODS OUTPUT statements.

```
ODS OUTPUT
  BRESLOWESTIMATES=&ds._breslow;
PROC LIFETEST <...>
ODS OUTPUT CLOSE;
```

Cartesian Product with PROC SQL

As described, for the estimation of an age and sex identical reference population to the population under risk for all event time points, the Cartesian product of all individuals (patid) and event time points is to be estimated and the created amount of data restricted to

the population at risk at the appropriate point in time. In the following PROC SQL program step, both requirements will be implemented: the restriction occurs in the WHERE condition, the rest of the SQL syntax is used for the creation of the Cartesian product. A prerequisite is, besides the output data set, another data set which includes only the event time point per strata without repetitions. This is achieved with the PROC FREQ.

```
PROC FREQ DATA=&ds;
  TABLES time*&strat /NOPRINT
  OUT=&ds._ereig(KEEP=time &strat
                 RENAME=(time=globtime
                        &strat=globstrat));
  WHERE status=1;
RUN;
PROC SQL NOPRINT;
  CREATE TABLE kart AS
  SELECT a.patid, a.time, a.&strat,
         b.globtime, b.globstrat
  FROM &ds a, &ds._ereig b
  WHERE (a.time>=b.globtime AND
         a.&strat=b.globstrat)
  ORDER BY a.patid, b.globtime;
QUIT;
```

ACKNOWLEDGMENT

The authors thank Anne Neumann for English translation and formatting of the manuscript.

REFERENCES

- Aalen OO. 1978. Nonparametric inference for a family of counting processes. *The Annals of Statistics* 6 (4): 701-726.
- Allison P. 2010. Survival analysis using SAS - a practical guide (2nd ed.). Cary, NC: SAS Publishing.
- Borgan O, and Liestol K. 1990. A Note on Confidence Interval and Bands for the Survival data: theory, empirical evaluation, computational realization and applications. *Scand J Statistics* 18: 35-41.
- Brenner H, and Gefeller O. 1996. An alternative approach to monitoring cancer patient survival. *Cancer* 78 (9): 2004-2010.
- Brenner H, Gefeller O, and Hakulinen T. 2004. Period analysis for 'up-to-date' cancer survival data: theory, empirical evaluation, computational realisation and applications. *Eur. J Cancer* 40 (3): 326-335.
- Colosimo EA, Ferreira FF, Oliveira MD, and Sousa CB. 2002. Empirical comparisons between Kaplan-Meier and Nelson-AAalen survival function estimators. *J Stat Comput Simul* 72 (4): 299-308.
- Cutler SJ, and Ederer F. 1958. Maximum utilization of the life table method in analyzing survival. *J Chronic. Dis.* 8 (6): 699-712.
- Dickman P. 2004. Estimating and modelling relative survival using SAS, manual.
- Ederer F, and Heise H. 1959. Instructions to IBM 650 programmers in processing survival computations - methodological note No. 10 - end results section. Bethesda, MD: National Cancer Institute.
- Geiss K, Meyer M, Radespiel-Troger M, and Gefeller O. 2009. SURVSOFTE-Software for nonparametric survival analysis. *Comput Methods Programs Biomed.* 96 (1): 63-71.
- Hakulinen T. 1982. Cancer survival corrected for heterogeneity in patient withdrawal. *Biometrics* 38 (4): 933-942.
- Kalbfleisch JD, and Prentice RL. 1980. *Statistical analysis of failure time data*. New York: Wiley.
- Koch R. 2001. Analyse von Überlebensdauervertellungen unter Berücksichtigung der Altersstruktur der Patienten in Klinischen Krebsregistern. In *Jahresbericht 2001 des Tumorzentrums Dresden e.V.*, 67-74.
- Krämer W, Schoffer O, and Tschiersch L. 2008. *Datenanalyse mit SAS*. Heidelberg: Springer.
- Pohar Perme M, Stare J, and Estève J. 2012. On Estimation in Relative Survival. *Biometrics* 68: 112-120.
- SAS Institute Inc. 2009. SAS 9.2 Macro Language - Reference. Cary, NC: SAS Publishing.
- SAS Institute Inc. 2010. SAS/STAT 9.22 User's Guide. Cary, NC: SAS Publishing.
- Statistisches Bundesamt. 2011. Periodensterbetafeln für Deutschland 1871/1881 bis 2007/2009. Wiesbaden: Statistisches Bundesamt.
- Wittekind C, and International Union Against Cancer (UICC). 2003. *TNM-Klassifikation maligner Tumoren*. Berlin ; Heidelberg [u.a.]: Springer.